

## Aims

This exercise aims to get you to practice:

- AWS EC2
- AWS S3
- Hadoop MapReduce on AWS EMR

## Background

### AWS EC2:

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage. See more documentation at:

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html>.

### AWS S3:

Amazon Simple Storage Service (Amazon S3) is storage for the Internet. You can use Amazon S3 to store and retrieve any amount of data at any time, from anywhere on the web. You can accomplish these tasks using the AWS Management Console, which is a simple and intuitive web interface. See more documentation at:

<http://docs.aws.amazon.com/AmazonS3/latest/gsg/GetStartedWithS3.html>

### AWS EMR:

Amazon EMR is a web service that makes it easy to quickly and cost-effectively process vast amounts of data. Amazon EMR simplifies big data processing, providing a managed Hadoop framework that makes it easy, fast, and cost-effective for you to distribute and process vast amounts of your data across dynamically scalable Amazon EC2 instances. You can also run other popular distributed frameworks such as Apache Spark in Amazon EMR, and interact with data in other AWS data stores such as Amazon S3. See more documentation at:

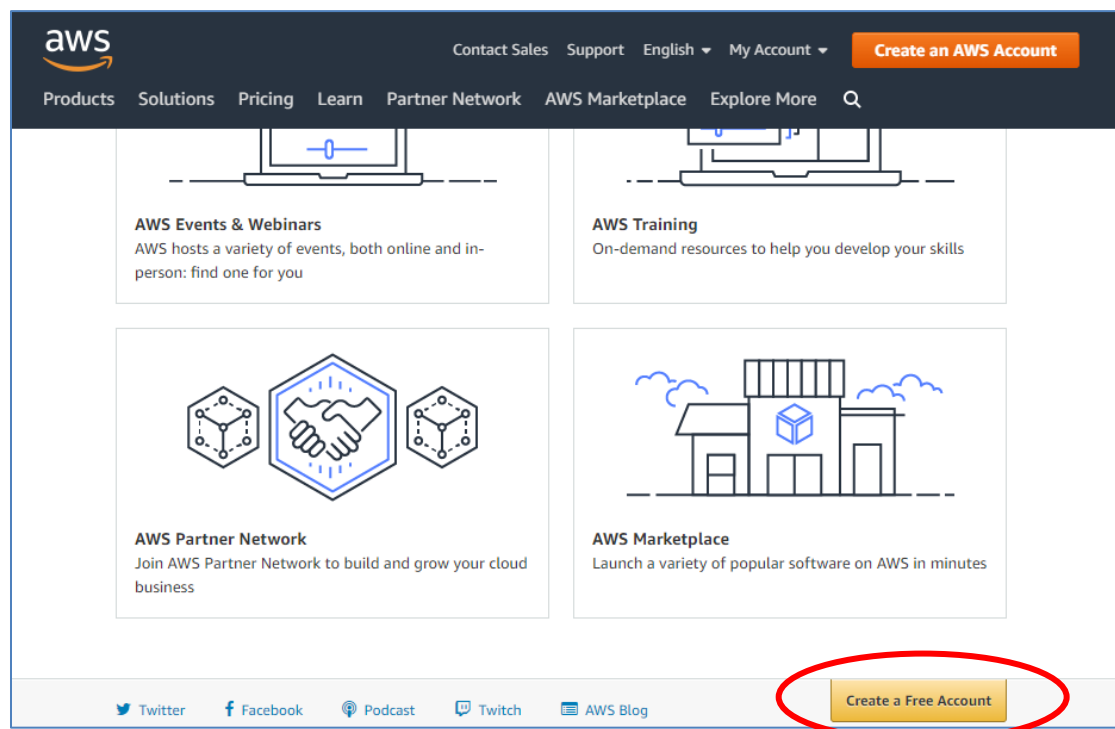
[https://docs.aws.amazon.com/emr/index.html#lang/en\\_us](https://docs.aws.amazon.com/emr/index.html#lang/en_us)

**Caution: Before doing the lab, please make sure that you have redeemed your \$100 credits!!! We are NOT responsible for any charge of your credit cards if you do not follow the lab instructions.**

# Register AWS

If you have an existing Amazon.com account (which you use for shopping on Amazon.com), you can use the same email and password for AWS. Otherwise, please follow the below instructions:


- Go to <http://aws.amazon.com> and click “Create a Free Account”
- Enter your contact information, credit card details, and do a phone verification. The whole process will take ~5 minutes.
- Now you can login to the AWS console at <http://console.aws.amazon.com> with your credentials.




- After logging into AWS, you can obtain your AWS account ID at <https://console.aws.amazon.com/billing/home?#/account>

## Apply for AWS Educate


- Go to <https://aws.amazon.com/education/awseducate/apply/>
- Click “Apply for AWS Educate for students”
- Provide all the information as required, using your UNSW email to verify.
- You may need to wait for several minutes to receive the confirmation of your application in your UNSW email, which contains the promo code.
- Please select “Enter an AWS Account Id”. Use the account Id you obtained in the previous step.

 **Click here to enter an AWS Account ID**  
 Approved students are sent a welcome email and benefits including an AWS promotional code.

[Don't have one? Sign up now](#)







 **Click here to select an AWS Educate Starter Account**  
 An AWS Educate Starter Account is a free, capped-account that doesn't require a credit card. There are some Usage limitations including an approximately 25% reduction in access to AWS services. Because Starter Accounts are capped, a separate AWS promotional code is not provided.  
[Frequently Asked Questions](#)


Please note that any personal information you provide will be treated in accordance with the [AWS Educate Terms and Conditions](#) and [AWS Privacy Notice](#)

**NEXT** 

## Redeem your credits


- Sign into your account at <https://console.aws.amazon.com>.
- In the upper right corner, click on the arrow next to your name and go to Billing & Cost Management.
- Next, in your Dashboard menu on the left, click on Credits and once you are there, you will be able to see all the relevant info such as the remaining balance, applicable products and services, and expiration date.
- Enter the credit code and the captcha, and you should be done. You should see a table appear which shows how many credits you have left.
- The last column has a link “See complete list” which lists the AWS products supported with the credit code. The credits cover all AWS products that you may need in your project. **If you use anything not on this list, your credit card will be charged (!!!)**


 **AWS**  **Services**  **Edit**  cxsyx Global  Support 

**Credits** 

Please enter your code below to redeem your credits.

Promo Code

Security Check 

 [Refresh Image](#)

Please type the characters as shown above

By clicking "Redeem" you indicate that you have read and agree to the terms of the AWS Promotional Credit Terms & Conditions located [here](#).

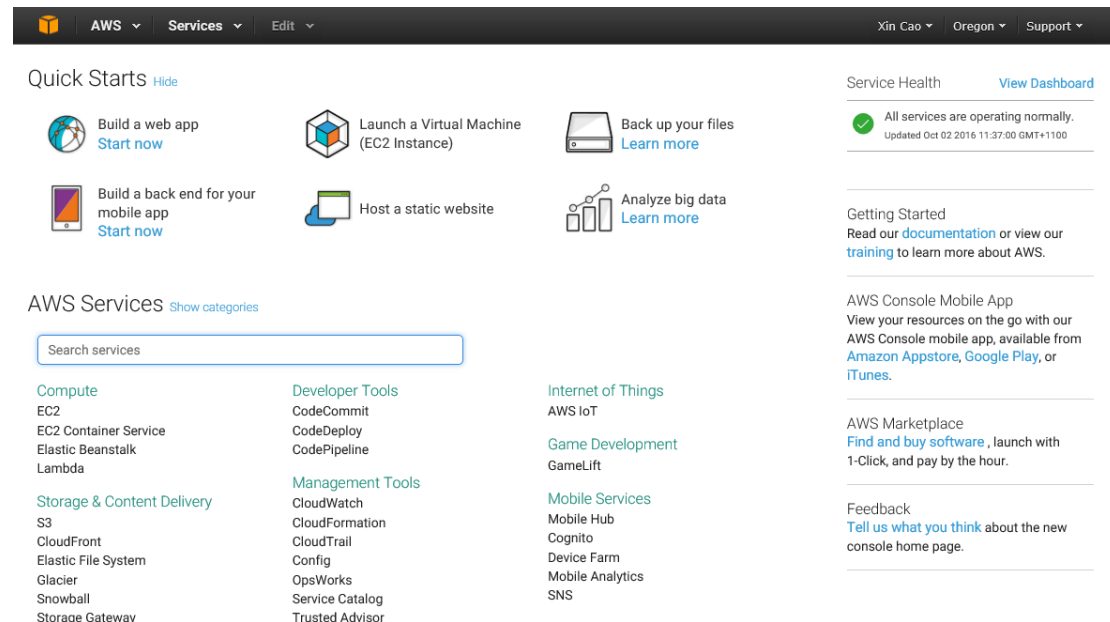
**Redeem**

Below are all the credits you have redeemed with AWS. Credits will automatically be applied to your bill. Only credits that apply to a specific service can be used.

Expiration Date	Credit Name	Credits Used	Credits Remaining	Applicable Products
2017-05-31	ENG_FY2016_Q2_05_100USD	\$0.00	\$100.00	<a href="#">See complete list</a>

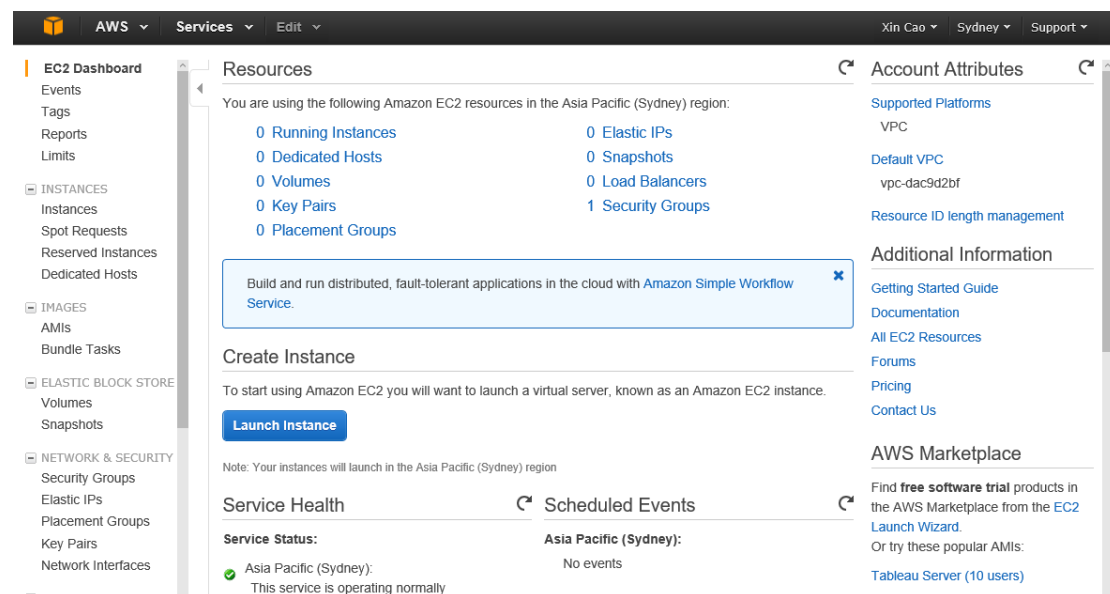
# Try AWS EC2 Using Free Tier Accounts

1. Log in AWS using your own account. Once you have signed in, you will be greeted by a page like this:



**Caution: Make sure that the region information on the top right is set to “Sydney”. If it is not, change it to Sydney by selecting from the dropdown menu there.**

2. Click on the EC2 link (first link under the Compute category). You will go to a dashboard page like this:



3. Click the blue “Launch Instance” button, and you will be redirected to a page like the following:

**Step 1: Choose an Amazon Machine Image (AMI)**

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.

**Quick Start**

- My AMIs
- AWS Marketplace
- Community AMIs
- ☐ Free tier only ⓘ

AMI	Description	Root device type	Virtualization type	Architecture
<b>Amazon Linux AMI 2016.09.0 (HVM), SSD Volume Type</b> - ami-55d4e436	The Amazon Linux AMI is an EBS-backed, AWS-supported image. The default image includes AWS command line tools, Python, Ruby, Perl, and Java. The repositories include Docker, PHP, MySQL, PostgreSQL, and other packages.	ebs	hvm	64-bit
<b>Red Hat Enterprise Linux 7.2 (HVM), SSD Volume Type</b> - ami-e0c19f83	Red Hat Enterprise Linux version 7.2 (HVM), EBS General Purpose (SSD) Volume Type	ebs	hvm	64-bit
<b>SUSE Linux Enterprise Server 12 SP 1 (HVM), SSD Volume Type</b> - ami-0f510a6c	SUSE Linux Enterprise Server 12 Service Pack 1 (HVM), EBS General Purpose (SSD) Volume Type. Public Cloud, Advanced Systems Management, Web and Scripting, and Legacy modules enabled.	ebs	hvm	64-bit
<b>Ubuntu Server 14.04 LTS (HVM), SSD Volume Type</b> - ami-ba3e14d9	Ubuntu Server 14.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical ( <a href="http://www.ubuntu.com/cloud/services">http://www.ubuntu.com/cloud/services</a> ).	ebs	hvm	64-bit

You can use many AMIs (Amazon Machine Image) to finish your task. In this lab, we will use the Ubuntu AMI, and continue to the next step to choose your instance type.

4. Choose the instance type t2.micro, and click on “Review and Launch”.

**Caution: This is the only one that is free tier eligible. You will be billed if you select other instance types!!**

**Step 2: Choose an Instance Type**

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. [Learn more](#) about instance types and how they can meet your computing needs.

**Filter by:** All instance types | Current generation | [Show/Hide Columns](#)

**Currently selected:** t2.micro (Variable ECUs, 1 vCPUs, 2.5 GHz, Intel Xeon Family, 1 GiB memory, EBS only)

	Family	Type	vCPUs ⓘ	Memory (GiB)	Instance Storage (GB) ⓘ	EBS-Optimized Available ⓘ	Network Performance ⓘ
<input type="checkbox"/>	General purpose	t2.nano	1	0.5	EBS only	-	Low to Moderate
<input checked="" type="checkbox"/>	General purpose	t2.micro	1	1	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.small	1	2	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.medium	2	4	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.large	2	8	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	m4.large	2	8	EBS only	Yes	Moderate

[Cancel](#) [Previous](#) [Review and Launch](#) [Next: Configure Instance Details](#)

5. In the next page, click on Launch.

**Step 7: Review Instance Launch**

Please review your instance launch details. You can go back to edit changes for each section. Click **Launch** to assign a key pair to your instance and complete the launch process.

**⚠ Improve your instances' security. Your security group, launch-wizard-1, is open to the world.**

Your instances may be accessible from any IP address. We recommend that you update your security group rules to allow access from known IP addresses only. You can also open additional ports in your security group to facilitate access to the application or service you're running, e.g., HTTP (80) for web servers. [Edit security groups](#)

▼ **AMI Details** [Edit AMI](#)

**Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-ba3e14d9**

Free tier eligible Ubuntu Server 14.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (<http://www.ubuntu.com/cloud/services>).  
Root Device Type: ebs Virtualization type: hvm

▼ **Instance Type** [Edit instance type](#)

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
t2.micro	Variable	1	1	EBS only	-	Low to Moderate

▼ **Security Groups** [Edit security groups](#)

[Cancel](#) [Previous](#) [Launch](#)

6. You will be then prompted to create or use an existing key-pair. Create a new one by choosing “Create a new key pair” from the drop-down menu and giving it some name of your choice (e.g., “comp9313”). You should then download the key pair, and keep it somewhere that you won’t accidentally delete. **Remember that there is NO WAY to get to your instance if you lose your key.**

**Caution: Don't select the Proceed without a key pair option. If you launch your instance without a key pair, then you can't connect to it.**

**Step 7: Review Instance Launch**

Please review your instance launch details. You can go back to edit changes for each section. Click **Launch** to assign a key pair to your instance and complete the launch process.

**⚠ Improve your instances' security. Your security group, launch-wizard-1, is open to the world.**

Your instances may be accessible from any IP address. We recommend that you update your security group rules to allow access from known IP addresses only. You can also open additional ports in your security group to facilitate access to the application or service you're running, e.g., HTTP (80) for web servers. [Edit security groups](#)

▼ **AMI Details** [Edit AMI](#)

**Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-ba3e14d9**

Free tier eligible Ubuntu Server 14.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (<http://www.ubuntu.com/cloud/services>).  
Root Device Type: ebs Virtualization type: hvm

▼ **Instance Type** [Edit instance type](#)

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
t2.micro	Variable	1	1	EBS only	-	Low to Moderate

▼ **Security Groups** [Edit security groups](#)

[Cancel](#) [Previous](#) [Launch](#)

**Select an existing key pair or create a new key pair**

A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about [removing existing key pairs from a public AMI](#).

Create a new key pair

**Key pair name**

comp9313

[Download Key Pair](#)

**ⓘ You have to download the private key file (\*.pem file) before you can continue. Store it in a secure and accessible location. You will not be able to download the file again after it's created.**

[Cancel](#) [Launch Instances](#)

7. Once you download your key, you should change the permissions of the key to user-only RW. Move the file to your home folder, and then do:

```
$ chmod 600 comp9313.pem
```

8. After this is done, click on “Launch Instances”, and you should see a screen showing that your instances are launching:

The screenshot shows the AWS Management Console 'Launch Status' page. At the top, there's a navigation bar with 'AWS', 'Services', 'Edit', and user information 'cxsyzx', 'Sydney', and 'Support'. Below the navigation bar, the page title is 'Launch Status'. A green notification box states: 'Your instances are now launching. The following instance launches have been initiated: i-02d2c60e60a29749a. View launch log'. Below this, a blue information box says: 'Get notified of estimated charges. Create billing alerts to get an email notification when estimated charges on your AWS bill exceed an amount you define (for example, if you exceed the free usage tier)'. A section titled 'How to connect to your instances' explains that instances are launching and will be in the 'running' state. It also mentions clicking 'View Instances' to monitor status. At the bottom, there's a footer with 'Feedback', 'English', copyright '© 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved.', 'Privacy Policy', and 'Terms of Use'.

9. Click on “View Instances” to see your instance state. It should change to “Running” and “2/2 status checks passed” as shown below within some time. You are now ready to ssh into the instance.

The screenshot shows the AWS Management Console 'View Instances' page. The left sidebar has a navigation menu with 'EC2 Dashboard', 'Events', 'Tags', 'Reports', 'Limits', 'INSTANCES', 'Images', 'Elastic Block Store', and 'Snapshots'. The main content area has tabs for 'Launch Instance', 'Connect', and 'Actions'. Below these is a search bar and a table of instances. The table has columns: Name, Instance ID, Instance Type, Availability Zone, Instance State, and Status. One instance is listed: 'i-02d2c60e60a29749a', 't2.micro', 'ap-southeast-2c', 'running'. Below the table, there's a detailed view for the selected instance. It shows the 'Description' tab, 'Status Checks' (2/2 passed), 'Monitoring', and 'Tags'. The 'Status Checks' section shows 'Instance state: running'. The 'Public IP' is highlighted with a red box and is '52.64.199.38'. Other details include 'Public DNS: ec2-52-64-199-38.ap-southeast-2.compute.amazonaws.com', 'Private DNS: ip-172-31-10-167.ap-southeast-2.compute.amazonaws.com', 'Elastic IPs', and 'Availability zone: ap-southeast-2c'. The footer is the same as the previous screenshot.

10. Note down the Public IP of the instance from the instance listing (in the example, it is 52.64.199.38). Then, do:

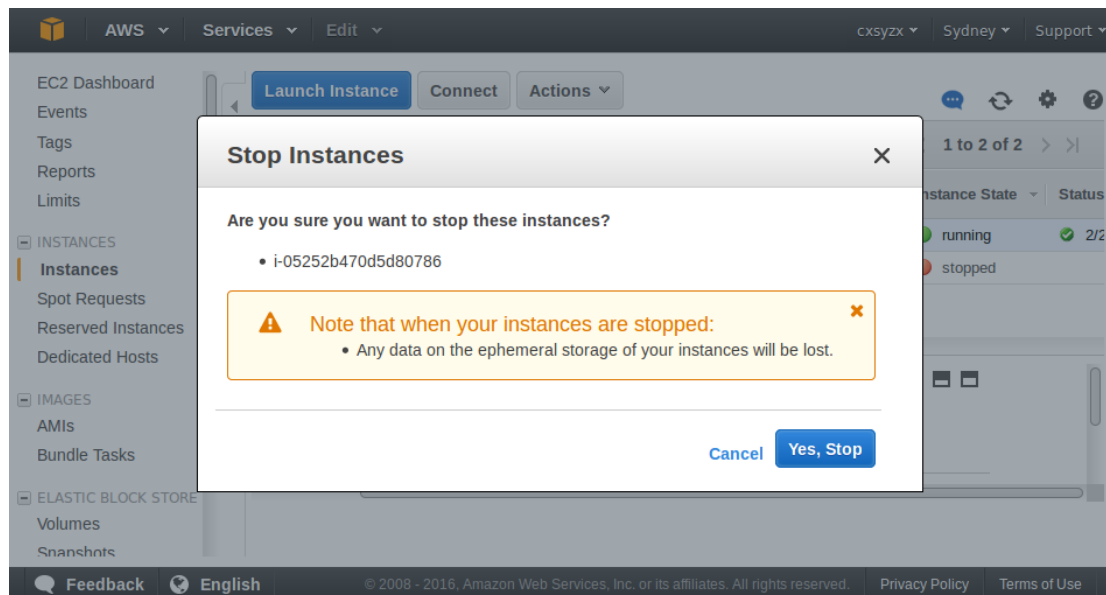
```
$ ssh -i ~/comp9313.pem ubuntu@52.64.199.38
```

Alternatively, you can also use the public DNS to connect to the instance.

If everything works fine, you should be able to ssh to the AWS instance.

11. To shut down the instance, right click the instance and select “Instance State -> Stop”. Then confirm to stop the instance.

**Caution: If you choose “terminate”, then all the files in this instance will be lost permanently, and you cannot use it again!**



12. You can also launch another instance. This time, after the step “Review and Launch”, click “Edit security groups” (a security group is a set of firewall rules that control the traffic for your instance).

▼ Security Groups [Edit security groups](#)

Security group name	launch-wizard-2
Description	launch-wizard-2 created 2016-10-03T04:38:25.934+11:00

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ
SSH	TCP	22	0.0.0.0/0

Then, choose the existing security group you created for the first instance.

AWS Management Console - Step 6: Configure Security Group

Assign a security group: ☐ Create a new security group ☒ Select an existing security group

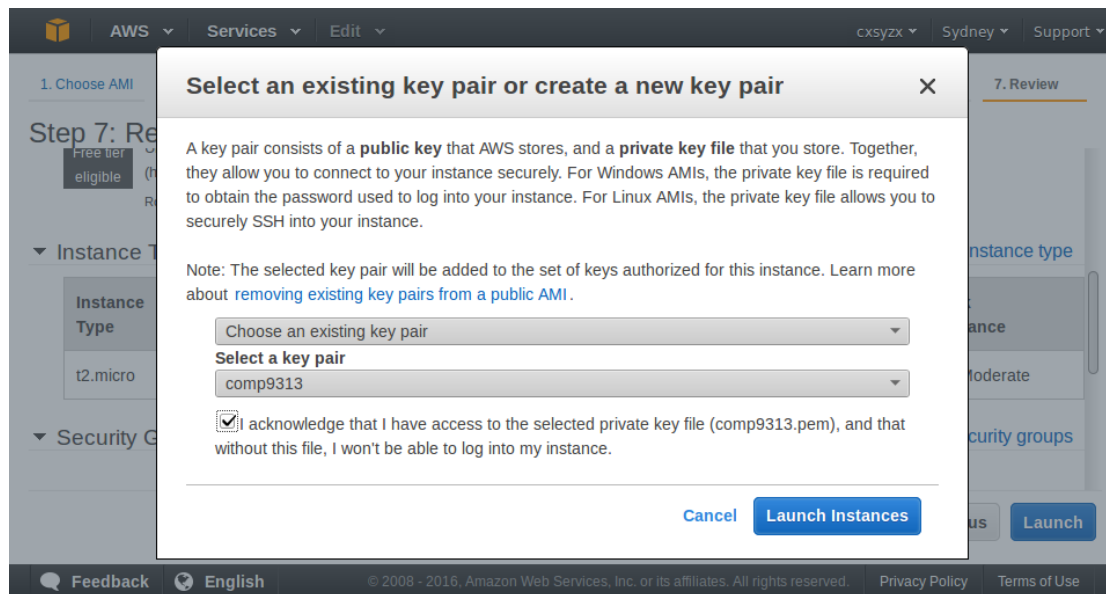
Security Group ID	Name	Description	Actions
<input type="checkbox"/> sg-c3133aa7	default	default VPC security group	<a href="#">Copy to new</a>
<input checked="" type="checkbox"/> sg-be9cb5da	launch-wizard-1	launch-wizard-1 created 2016-10-03T04:32:26.947+11:00	<a href="#">Copy to new</a>

Inbound rules for sg-be9cb5da (Selected security groups: sg-be9cb5da)

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ
SSH	TCP	22	0.0.0.0/0



Next, you can use your existing key pair to launch the instance.

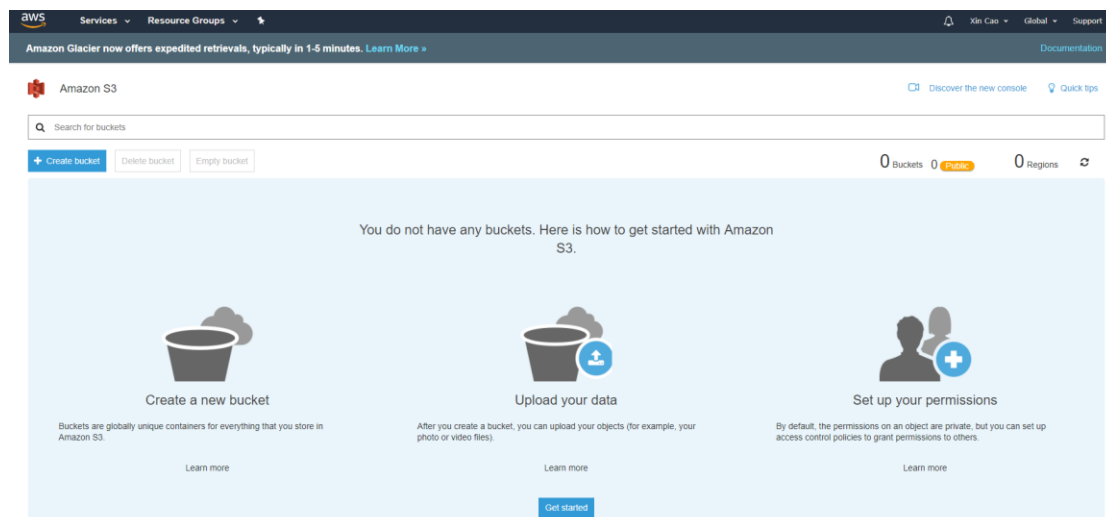


**Caution: You will be billed for AWS instances as they are alive, so you will want to terminate them when they aren't in direct use! Here are the Amazon instructions. Always remember to terminate the instances if they will not be used any more. You can stop an instance if you still need to use it later.**

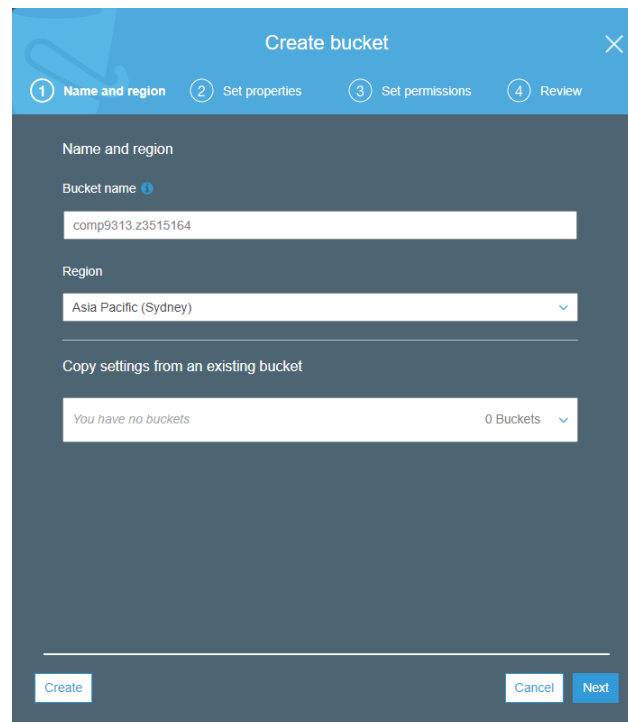
## Store Data in AWS S3

### Create a Bucket in S3

1. Every object in Amazon S3 is stored in a bucket (like a folder in your local file system). Before you can store data in Amazon S3 you must create a bucket. Go back to the AWS Management Console and open the Amazon S3 console.



2. Click Create Bucket. The “Create bucket” dialog box appears.



Enter a bucket name in the Bucket Name field. The bucket name you choose must be unique across all existing bucket names in Amazon S3. For example, the tutorial names the bucket as “comp9313”.

Bucket names must comply with the following requirements:

- Can contain lowercase letters, numbers, periods (.) and dashes (-)
- Must start with a number or letter
- Must be between 3 and 255 characters long
- Must not be formatted as an IP address (e.g., 265.255.5.4)

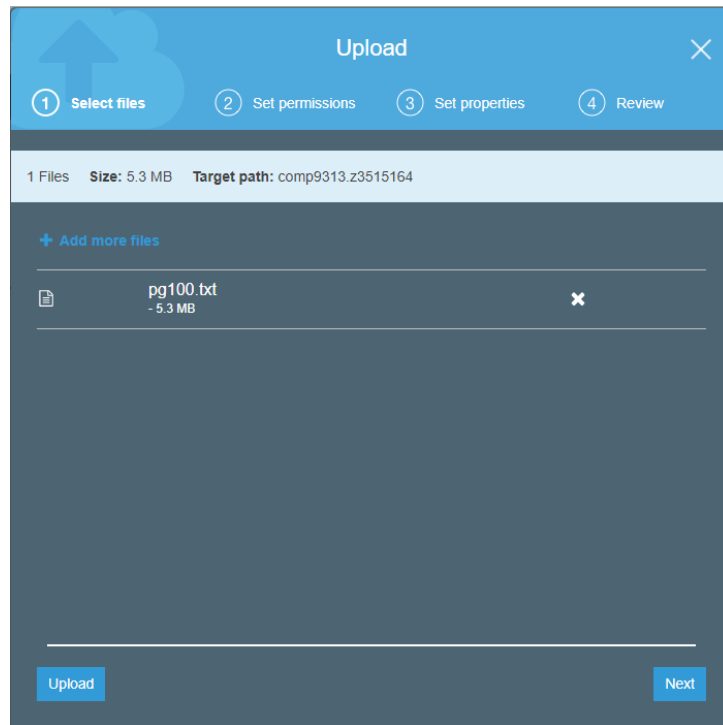
**Caution:** Because S3 allows your bucket to be used as a URL that can be accessed publicly, the bucket name that you choose must be globally unique. If some other account has already created a bucket with the name that you chose, you must use another name. Therefore, it is recommended to name your bucket as “comp9313.<YOUR\_StudentID>” (e.g., comp9313.z3515164).

In the Region drop-down list box, select region “Asia Pacific (Sydney)”, and click “Create”. You would see a bucket created for you in S3.

### **Add and Manage Files in a Bucket:**

Now that you've created a bucket, you're ready to add an object to it. An object can be any kind of file: a text file, a photo, a video and so forth.

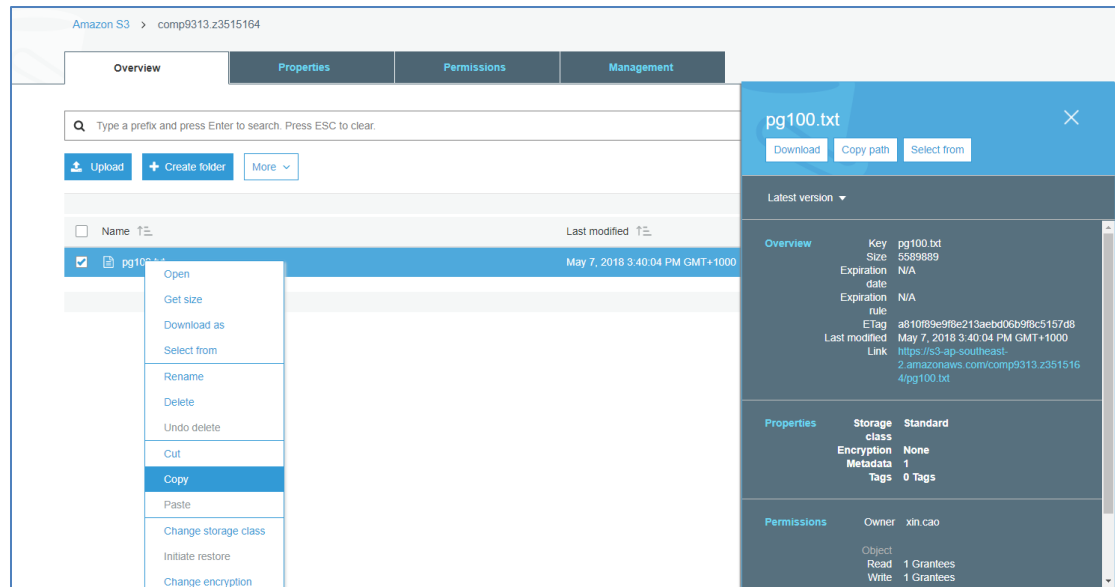
When you add a file to Amazon S3, you have the option of including metadata with the file and setting permissions to control access to the file.



In the Amazon S3 console click the bucket you want to upload an object into and then click “Upload”. The Upload dialog box then opens (appearance may differ slightly in different browsers). Download the pg100.txt file, create a folder “input” in your bucket, and upload it into the folder.

**Caution: The free tier account only has 5GB S3 storage. If your files exceed this space limit, you will be billed for the service!!!**

You can do various actions on the files in your bucket. Select the file to be managed, then click “Actions”, in the menu you can see all the actions you can do, such as Rename, Cut, and Copy. You can also view the properties of the file.



Finally, prepare a WordCount jar file, and upload it to AWS S3.

a) Download the WordCount.java used in Lab 3 from the course home page. Set the number of Reducers as 3. Compile the file and package the MapReduce program as a jar file wc.jar.

c) Test the jar file in your local machine first before uploading to S3.

## Run MapReduce Tasks on AWS EMR (Please first redeem the credits and then work on this!)

1. Go back to the AWS Management console and open the EMR console.



2. Choose Create cluster. On the Create Cluster page, you need to do the following:

### In General Configuration section:

a) Cluster name: comp9313.lab9

b) Logging: Select

By default, clusters created using the console have logging enabled. This option determines whether Amazon EMR writes detailed log data to Amazon S3.

When this value is set, Amazon EMR copies the log files from the EC2 instances in the cluster to Amazon S3. Logging to Amazon S3 can only be enabled when the cluster is created.

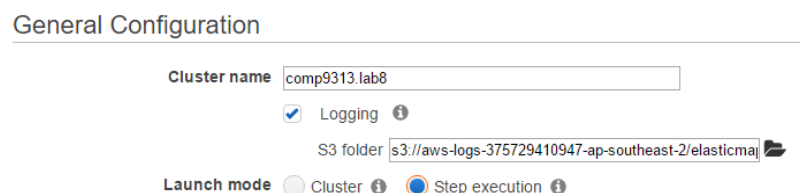
Logging to Amazon S3 prevents the log files from being lost when the cluster ends and the EC2 instances hosting the cluster are terminated. These logs are useful for troubleshooting purposes.

c) S3 folder: use default. The folder is used to store the logs.

You can also type or browse to your Amazon S3 bucket to store the Amazon EMR logs; for example, `s3://YOUR_BUCKET/logs`, or you can allow Amazon EMR to generate an Amazon S3 path for you. If you type the name of a folder that does not exist in the bucket, it is created for you.

d) Launch mode: select “Step execution.”

If you select “Cluster”, the instances will keep running after your MapReduce task is finished. However, you can do more jobs without creating a new cluster. By selecting “Step execution”, the instances will be terminated once the task is completed.



General Configuration

Cluster name

☒ Logging ⓘ

S3 folder

Launch mode ☐ Cluster ⓘ ☒ Step execution ⓘ

### In Add steps section:

a) Set the step type as Custom JAR

b) Click “Configure”, set Name as “WordCount”, set JAR location as “`s3://comp9313.<YOUR_StudentID>/wc.jar`”, set Arguments as “`comp9313.lab3.WordCount s3://comp9313.<YOUR_StudentID>/input s3://comp9313.<YOUR_StudentID>/output`”, select “Terminate cluster” for Action on Failure, and finally click Add.



AWS Services Edit cxyzx Sydney Support

Amazon EMR

Cluster list Security configurations VPC subnets Help

Cluster: comp9313.lab8 **Starting**

Connections: --  
Master public DNS: --  
Tags: -- [View All / Edit](#)

**Summary**

ID: j-1MVTVS9TE9G0  
Creation date: 2016-10-03 05:25 (UTC+11)  
Elapsed time: --  
Auto-terminate: No  
Termination protection: Off [Change](#)

**Configuration Details**

Release label: emr-5.0.0  
Hadoop distribution: Amazon 2.7.2  
Applications: Ganglia 3.7.2, Hive 2.1.0, Hue 3.10.0, Mahout 0.12.2, Pig 0.16.0, Tez 0.8.4  
Log URI: s3://aws-logs-375729410947-ap-southeast-2/elasticmapreduce/  
EMRFS Disabled

[Add step](#) [Resize](#) [Clone](#) [Terminate](#) [AWS CLI export](#)

Later, you will see the information for Connections and Master public DNS is updated, since the cluster is already started.

Click “Steps”, and you should see two jobs listed.

Steps

[Add step](#) [Clone step](#)

Filter: All steps 2 steps (all loaded)

ID	Name	Status	Start time (UTC+11)	Elapsed time	Log files	Actions
s-1LNGF60SCX88U	Setup hadoop debugging	Pending			<a href="#">View logs</a>	<a href="#">View jobs</a>
s-TVET1HK7UM3H	WordCount	Pending			<a href="#">View logs</a>	<a href="#">View jobs</a>

7. Wait until the WordCount task is finished. **Note that this may take several minutes.**

In the meantime, you can begin working on the next section, and go back to check the results later.

8. If the task is completed, you should see:

Cluster: comp9313.lab8 **Terminated** Steps completed

Connections: --  
Master public DNS: ec2-52-63-32-210.ap-southeast-2.compute.amazonaws.com [SSH](#)  
Tags: --

**Summary**

ID: j-F46WABPVF356  
Creation date: 2016-10-03 06:03 (UTC+11)  
End date: 2016-10-03 06:09 (UTC+11)  
Elapsed time: 6 minutes  
Auto-terminate: Yes  
Termination protection: Off

**Configuration Details**

Release label: emr-5.0.0  
Hadoop distribution: Amazon 2.7.2  
Applications: --  
Log URI: s3://aws-logs-375729410947-ap-southeast-2/elasticmapreduce/  
EMRFS consistent view: Disabled

**Network and Hardware**

Availability zone: ap-southeast-2b  
Subnet ID: subnet-1982495B  
Master: Terminated 1 m3.xlarge  
Core: Terminated 2 m3.xlarge  
Task: --

**Security and Access**

Key name: --  
EC2 instance profile: EMR\_EC2\_DefaultRole  
EMR role: EMR\_DefaultRole  
Visible to all users: All [Change](#)  
Security groups for sg-66b39a02 (ElasticMapReduce-Master: master)  
Security groups for sg-60b39a04 (ElasticMapReduce-Core & Task: slave)

**Monitoring**

**Hardware**

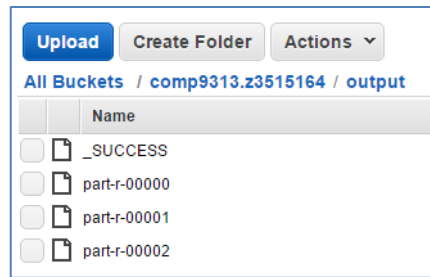
**Steps**

[Add step](#) [Clone step](#)

Filter: All steps 2 steps (all loaded)

ID	Name	Status	Start time (UTC+11)	Elapsed time	Log files	Actions
s-7QH2VHTCH1MD	WordCount	Completed	2016-10-03 06:07 (UTC+11)	42 seconds	<a href="#">View logs</a>	<a href="#">View jobs</a>
s-29PROAZEWF6R	Setup hadoop debugging	Completed	2016-10-03 06:07 (UTC+11)	2 seconds	<a href="#">View logs</a>	<a href="#">View jobs</a>

Go to your S3 bucket, the results should be stored there.



**Caution: Do not forget to terminate the cluster after you finish all labs!!! (Click “Terminate” and turn termination protection off)**